Universal Simulation with a Fidelity Criterion

Neri Merhav
Department of Electrical Engineering
Technion – I.I.T., Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

Marcelo J. Weinberger
Hewlett–Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304, U.S.A.
Email: marcelo@hpl.hp.com

***Abstract*— We consider the problem of universal simulation of memoryless sources and Markov sources, based on training sequence emitted from these sources. The objective is to maximize the conditional entropy of the simulated sequence given the training sequence, subject to a certain distance constraint between the probability distribution of the output sequence and the probability distribution of the input, training sequence. We derive a single–letter expression for the maximum conditional entropy and then propose a universal simulation scheme that asymptotically attains this maximum.**

I. Introduction

Simulation of a source means artificial production of random data with some probability law, by using a certain device that is fed by a source of purely random bits. Simulation of sources and channels is a problem that has been studied in a series of works, see, e.g., [1], [7], [8], [9] and references therein. In all these works, it was assumed that the probability law of the desired process is perfectly known.

Recently, a universal version of this problem was studied in [4], [5] (see also [2]), where the assumption of perfect knowledge of the target probability law was relaxed. Instead, the target source P to be simulated was assumed in [4] to belong to a certain parametric family $\mathcal{P}$, but is otherwise unknown, and a training sequence $X^m = (X_1, \ldots, X_m)$, that has emerged from this source, is available. In addition, the simulator is provided with a sequence of ℓ random bits $U^\ell = (U_1, \ldots, U_\ell)$, which is independent of X^m. The goal of the simulation scheme in [4] was to generate an output sequence $Y^n = (Y_1, \ldots, Y_n)$, $n \leq m$, corresponding to the simulated process, such that $Y^n = \psi(X^m, U^\ell)$, where ψ is a deterministic function that does not depend on the unknown source P, and which satisfies the following two conditions: (i) the probability distribution of Y^n is *exactly* the n-dimensional marginal of the probability law P corresponding to X^m for all $P \in \mathcal{P}$, and (ii) the mutual information $I(X^m; Y^n)$ is as small as possible, or equivalently (under (i)), the conditional entropy $H(Y^n|X^m)$ is as large as possible, simultaneously for all $P \in \mathcal{P}$ (so as to make the generated sample path Y^n as "original" as possible). In [4], the smallest achievable value of the mutual information (or, the largest conditional entropy) was characterized, and simulation schemes that asymptotically achieve these bounds were presented (see also [5]). In [3], the same simulation problem was studied in the regime of a delay–limited sytem, in which the simulator produces output samples on–line, as the training data is fed into the system sequentially. The cost of limited delay was characterized and a strictly optimum simulation system was proposed. A different perspective on universal simulation was investigated in [6], where x^m was assumed to be an individual sequence not originating from any probabilistic source.

In this work, we extend the scope of the universal simulation problem in another direction, namely, relaxing the requirement of *exact* preservation of the probability law at the output of the simulator. In particular, we study the best achievable tradeoff between the performance of the simulation scheme and the distance (measured in terms of a certain metric) between the probability law of the output and that of the input. Observe that when the probability law of the simulated sequence is not constrained to be identical to that of the training sequence, the criteria $\min I(X^m; Y^n)$ and $\max H(Y^n|X^m)$ are no longer equivalent. They both remain, however, reasonable measures of the "diversity" or the "richness" of the typical sample paths generated by the simulator. While the former criterion has been discussed in [5] (in the context of the $\bar{\rho}$–distance between probability distributions), here we focus on the latter.

For the class of discrete memoryless sources (DMSs), we derive a single-letter formula for the maximum achievable conditional entropy subject to the distance constraint, and propose a simulation scheme that universally achieves this performance for large m and n. We also briefly discuss how our derivations can be extended to the Markov case. Finally, we derive similar results for the $\bar{\rho}$–distance measure, which is not a special case of the distance measure considered in the first part.

II. Notation and Problem Formulation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets, as well as some other sets, will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets, will be denoted, respectively, by capital letters, the corresponding lower case letters, and calligraphic letters, all superscripted by their dimensions. For example, the random vector $X^m = (X_1, \ldots, X_m)$, (m – positive integer) may take a specific vector value $x^m = (x_1, \ldots, x_m)$ in $\mathcal{A}^m$, the mth order Cartesian power of $\mathcal{A}$, which is the alphabet of each component of this vector. For $i \leq j$ (i, j – integers), x_i^j will denote the segment $(x_i, \ldots, x_j)$, where for $i = 1$ the subscript will be omitted.

Let $\mathcal{P}$ denote the class of all DMSs with a finite alphabet $\mathcal{A}$, and let P denote a particular member of $\mathcal{P}$. For a given positive integer m, let $X^m = (X_1, X_2, \ldots, X_m)$, $X_i \in \mathcal{A}$, $i = 1, \ldots, m$, denote an m-vector drawn from P, namely, $\Pr\{X_i = x_i,\ i = 1, \ldots, m\} = \prod_{i=1}^m P(x_i) \stackrel{\Delta}{=} P(x^m)$ for every $(x_1, \ldots, x_m)$, $x_i \in \mathcal{A}$, $i = 1, \ldots, m$. Let $H \equiv H(X) = -\sum_{x\in\mathcal{A}} P(x) \log P(x)$ denote the entropy of the source P, where here and throughout the sequel $\log(\cdot) \stackrel{\Delta}{=} \log_2(\cdot)$. When it is the dependence of the entropy upon P that we wish to emphasize (rather than the name of the random variable X), we denote the entropy by $H(P)$, with a slight abuse of notation.

For given positive integers m, ℓ, and n, and for a given mapping $\psi : \mathcal{A}^m \times \{0,1\}^\ell \to \mathcal{A}^n$, let $Y^n = \psi(X^m, U^\ell)$. Let $W(y^n|x^m)$ denote the conditional probability of $Y^n = y^n$ given $X^m = x^m$ corresponding to the channel from X^m to Y^n that is induced by ψ. The expectation operator, denoted $\boldsymbol{E}\{\cdot\}$, will be understood to be taken with respect to (w.r.t.) the joint distribution $P \times W$ of (X^m, Y^n).

Let $\rho(P, Q)$ denote a distance measure between two probability measures on $\mathcal{A}$, and define the distance between P^n and Q^n, which are two probability measures on $\mathcal{A}^n$, as

$$\rho_n(P^n, Q^n) = \frac{1}{n}\sum_{i=1}^n \sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot|a^{i-1}), Q(\cdot|a^{i-1})).$$

For example, if $\rho(P(\cdot|a^{i-1}), Q(\cdot|a^{i-1}))$ is $\sum_{a_i} Q(a_i|a^{i-1}) \log[Q(a_i|a^{i-1})/P(a_i|a^{i-1})]$, then ρ_n is the normalized divergence between Q^n and P^n. In that sense, ρ_n can be thought of as a generalized divergence.[1]

Finally, let $H(Y^n|X^m)$ denote the conditional entropy of Y^n given X^m that is induced by the source P and the channel W (or, equivalently, the mapping ψ).

This paper is about the quest for a mapping ψ that is independent of the unknown P, and that satisfies the following conditions:

C1. For every $P \in \mathcal{P}$, the probability distribution Q^n of $Y^n = \psi(X^m, U^\ell)$ obeys $\rho_n(P^n, Q^n) \le D$, where P^n is the n–th power of P (i.e., the product measure corresponding to the DMS P, generating n–tuples), and D is a prescribed constant. Note that Q^n need not be necessarily memoryless.

C2. The mapping ψ maximizes $H(Y^n|X^m)$ simultaneously for all $P \in \mathcal{P}$ among all mappings satisfying C1.

III. Main Result

Let us define the function:

$$\phi(D) = \max\{H(Q) :\ \rho(P, Q) \le D\}, \tag{1}$$

and $\bar{\phi}(D) = \mathrm{UCE}\{\phi(D)\}$, where UCE stands for upper concave envelope. Note that if $\rho(P, \cdot)$ is convex in Q (which is the case for many useful metrics), then ϕ is concave, thus $\bar{\phi}(D) \equiv \phi(D)$. Our first theorem asserts that $\bar{\phi}(D)$ is an upper bound on the conditional entropy per symbol for any simulation scheme.

Theorem 1: (Converse): For every simulation scheme ψ that satisfies condition C1, $H(Y^n|X^m) \le n\bar{\phi}(D)$.

Discussion: (i) In fact, we prove below, moreover, that $H(Y^n) \le n\bar{\phi}(D)$. Intuitively, since the conditioning on X^m will be made (in the direct part, cf. Theorem 2 below) only via its empirical distribution, this conditioning does not make a big difference. (ii) Another obvious upper bound to $H(Y^n|X^m)$ is $\ell = nR$, where R is the key rate in bits per output symbol. However, if $R \le \bar{\phi}(D)$, then it makes sense to decrease D to the level that gives $\bar{\phi}(D) = R$, because larger values of D mean degrading the fidelity of the output distribution w.r.t. P, without any gain in the conditional entropy of the output. Thus, it can be assumed without loss of generality that $R \ge \bar{\phi}(D)$, i.e., the key–rate limitation is not really an issue. Moreover, by the same rationale, it makes sense to assume that $R \ge H(P)$, as otherwise, if $R < H(P)$, there is no incentive to allow $D > 0$, because then $H(Y^n|X^m)/n \le R < H(P) = \bar{\phi}(0)$, and so there is nothing to gain from distorting the probability law (this takes us back to the case $D = 0$). This means that the interesting situation occurs when the key rate is sufficiently large, and for the sake of simplicity, we will assume that it is unlimited, and focus only on the interplay between conditional entropy and fidelity.

Proof. Consider first the conditional entropy of the ith output symbol, Y_i, given Y^{i-1}. Then, we have:

$$\begin{aligned}
H(Y_i|Y^{i-1}) &= \sum_{a^{i-1}} Q(a^{i-1}) H(Q(\cdot|a^{i-1})) \\
&\le \sum_{a^{i-1}} Q(a^{i-1}) \phi(\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&\le \sum_{a^{i-1}} Q(a^{i-1}) \bar{\phi}(\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&\le \bar{\phi}\left(\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right).
\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}
&\frac{1}{n} H(Y^n|X^m) \\
\le\ & \frac{1}{n}\sum_{i=1}^n H(Y_i|Y^{i-1}) \\
\le\ & \frac{1}{n}\sum_{i=1}^n \bar{\phi}\left(\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right) \\
\le\ & \bar{\phi}\left(\frac{1}{n}\sum_{i=1}^n \sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1}))\right) \\
=\ & \bar{\phi}(\rho_n(P^n, Q^n)) \\
\le\ & \bar{\phi}(D),
\end{aligned} \tag{2}$$

which completes the proof of Theorem 1.

Theorem 2: (Direct): Assume that $\rho(P, Q)$ is: (i) continuous at P uniformly in Q, and (ii) continuous and bounded in

[1] In general, additive distance functions between the conditional distributions $\{P(\cdot|a^{i-1})\}$ and $\{Q(\cdot|a^{i-1})\}$ may arise naturally in prediction and sequential decision problems, as they reflect the penalty for mismatch between the assumed probability law and the underying one.

Q for a given P. Then, there exists a sequence of simulation schemes, independent of P, that asymptotically (as $m, n \to \infty$) satisfy condition C1, and whose conditional entropies tend to $n\bar{\phi}(D)$ for all $P \in \mathcal{P}$.

Our proposed universal simulation scheme (see proof below) is based on forming grids in $\mathcal{P}$ and 'quantizing' the empirical distribution of X^m to the nearest grid point, with the density of the grid growing slower than m. This will be needed to guarantee that the induced conditional distributions at the output would be close to Q^*, the achiever of $\phi(D)$ (cf. eqs. (6) and (7) below).

Sketch of Proof. We actually prove that $\phi(D)$ is achievable, which coincides with $\bar{\phi}(D)$ whenever ϕ is concave. If this is not the case, then time–sharing between two schemes should be applied, and the below description refers to the action to be carried out for each one of the two working points.

Let us form a sequence of grids, $\mathcal{P}_K = \{P_1, P_2, \ldots, P_K\}$, $K = 1, 2, \ldots$, such that $\cup_{K=1}^{\infty} \mathcal{P}_K$ is dense in the simplex of probability distributions over $\mathcal{A}$. For a given probability distribution P' on $\mathcal{A}$, let $[P']_K$ denote the[2] nearest neighbor of P' in $\mathcal{P}_K$ (under an arbitrary metric between probability distributions, which is not necessarily ρ, say, the variational distance). Thus, the distance between P' and $[P']_K$ is bounded uniformly by a number ϵ_K, which tends to zero as $K \to \infty$. Our simulation scheme works as follows: Given X^m, extract its empirical distribution, $\hat{P}$, and 'quantize' it to the nearest neighbor $\tilde{P} = [\hat{P}]_K \in \mathcal{P}_K$. Then, find the achiever $\tilde{Q}$ of $\phi(D)$ but with $\tilde{P}$ playing the role of P, and finally, use $\tilde{Q}$ as the target memoryless source that governs Y^n (which is implemented with unlimited key rate). Let $T_{\tilde{P}} = T_{[\hat{P}]_K}$ denote the union of all type classes $\{T_{x^m}\}$ for which $\tilde{P}$ is the nearest neighbor of the empirical distribution $\hat{P}$ corresponding to T_{x^m}. Now by the AEP, for any fixed K, the probability $P(T_{[P]_K})$ goes to unity as m grows without bound. Since ρ is continuous at P uniformly in Q, and $[P]_K$ is within distance ϵ_K from P, then $|\rho(P,Q) - \rho([P]_K, Q)| \le \delta_K$, where $\delta_K \to 0$ as $K \to \infty$, independently of Q. As for the conditional output entropy, we then have:

$$\begin{aligned}
\frac{1}{n}H(Y^n|X^m) &= \boldsymbol{E}\{H(\tilde{Q})\} \\
&\ge \sum_{T_{x^m} \subset T_{[P]_K}} P(T_{x^m}) H(\tilde{Q}) \\
&= \sum_{T_{x^m} \subset T_{[P]_K}} P(T_{x^m}) \times \\
&\quad \max\{H(Q):\ \rho([P]_K, Q) \le D\} \\
&\ge \sum_{T_{x^m} \subset T_{[P]_K}} P(T_{x^m}) \times \\
&\quad \max\{H(Q):\ \rho(P,Q) + \delta_K \le D\} \\
&= \sum_{T_{x^m} \subset T_{[P]_K}} P(T_{x^m}) \phi(D - \delta_K) \\
&= P(T_{[P]_K}) \phi(D - \delta_K). \qquad (3)
\end{aligned}$$

[2]We assume, without essential loss of generality, that there are no ties.

Since ϕ is concave, it is also continuous (except, perhaps for the edgepoints), and thus $\phi(D)$ is asymptotically achieved for large K and m.

It remains to show that $\rho(P^n, Q^n)$ is essentially less than D. Before we do that, we pause to introduce some additional notation, and a few facts that we will need in the sequel. Let the quantization of $\hat{P}$ result in $P_k = [\hat{P}]_K \in \mathcal{P}_K$, for some $k = 1, 2, \ldots, K$. Then, the corresponding achiever of $\phi(D)$, which we earlier denoted by $\tilde{Q}$, will also be denoted by Q_k. We will assume that $Q_1, \ldots, Q_K$ are all distinct (otherwise, we can slightly perturb some of them). We will also denote by k_0 the integer $k \in \{1, \ldots, K\}$ for which $P_k = [P]_K$. The corresponding Q_{k_0} will also be denoted by Q^*. For a given $\delta > 0$, let $\mathcal{T}_{Q_k}(\delta)$ denote the union of all $\{T_{x^m}\}$ corresponding to empirical distributions $\{\hat{P}\}$ for which $D(\hat{P}\|Q_k) \le \delta$. As $\{Q_k,\ k = 1, \ldots, K\}$ are assumed distinct, then there exists a small enough $\delta > 0$, such that $\mathcal{T}_{Q_k}(\delta)$ are disjoint. This follows from the fact that the divergence is lower bounded in terms of the variational distance, which is a metric. By the same token, it is easy to see that if l is sufficiently large and $\delta > 0$ is sufficiently small, and if $a^l \in \mathcal{T}_{Q_k}(\delta)$ for some k, then for any extension $a^{l+1} = (a^\ell, a_{l+1})$, the empirical distribution is still closer to Q_k (in the divergence sense) than to any $Q_{k'}$, $k' \ne k$.

Returning now to the proof that $\rho_n(P^n, Q^n)$ is not much larger than D, we will first show that for any $\epsilon > 0$ and sufficiently large n and m, $\rho(P(\cdot), Q(\cdot|a^{i-1}))$ is essentially less than D for all $i \ge \epsilon n$ and for all $a^{i-1} \in T_{Q^*}(\delta)$. To this end, let us examine the conditional distribution $Q(a_i|a^{i-1})$, induced by the proposed scheme, for $a^{i-1} \in T_{Q^*}(\delta)$.

$$\begin{aligned}
& Q(a_i|a^{i-1}) \\
&= \frac{\sum_{T_{x^m}} P(T_{x^m}) \tilde{Q}(a^i)}{\sum_{T_{x^m}} P(T_{x^m}) \tilde{Q}(a^{i-1})} \\
&= \frac{\sum_k P(T_{P_k}) Q_k(a^i)}{\sum_k P(T_{P_k}) Q_k(a^{i-1})} \\
&= \frac{P(T_{[P]_K}) Q^*(a^i) + \sum_{k \ne k_0} P(T_{P_k}) Q_k(a^i)}{P(T_{[P]_K}) Q^*(a^{i-1}) + \sum_{k \ne k_0} P(T_{P_k}) Q_k(a^{i-1})}. \qquad (4)
\end{aligned}$$

The first term in the numerator and the first term in the denominator are the desired terms. Let us assess the relative error contributed by each one of the other terms in the numerator and the denominator. As for the denominator, for every $k \ne k_0$, $P(T_{P_k}) \le 2^{-m\epsilon'_K}$ (for some $\epsilon'_K > 0$) and $Q_k(a^{i-1}) \le Q^*(a^{i-1})$ since $a^{i-1} \in \mathcal{T}_{Q^*}(\delta)$ and $i \ge n\epsilon$ (see the previous paragraph). The same goes for the numerator because, as explained earlier, the empirical distribution of a^i is still closer to Q^* than to any Q_k, $k \ne k_0$. Thus,

$$\begin{aligned}
Q(a_i|a^{i-1}) &\le \frac{P(T_{[P]_K}) Q^*(a^i)(1 + K \cdot 2^{-m\epsilon'_K})}{P(T_{[P]_K}) Q^*(a^{i-1})} \\
&= Q^*(a_i)(1 + K \cdot 2^{-m\epsilon'_K}) \qquad (5)
\end{aligned}$$

and by the same token, $Q(a_i|a^{i-1}) \ge Q^*(a_i)/(1 + K \cdot 2^{-m\epsilon'_K})$. Now, ρ is assumed continuous in Q. Thus, since we have just seen that $Q(\cdot|a^{i-1})$ is close to Q^* for large enough m and

i (for any metric), then $\rho(P, Q(\cdot|a^{i-1})) \le \rho(P, Q^*) + \mu_{m,K}$, where $\mu_{m,K} \to 0$ as $m \to \infty$ for every fixed K. Consider now the i–th term of the distance function ρ_n, where $i \ge \epsilon n$. Then,

$$\begin{aligned}
&\sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&= \sum_{T_{x^m}} P(T_{x^m}) \sum_{a^{i-1}} \tilde{Q}(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&= \sum_{k} P(T_{P_k}) \sum_{a^{i-1}} Q_k(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&= P(T_{[P]_K}) \sum_{a^{i-1}} Q^*(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) + \\
&\quad \sum_{k \ne k_0} P(T_{P_k}) \sum_{a^{i-1}} Q_k(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})), \quad (6)
\end{aligned}$$

where the second term vanishes as $P(T_{P_k})$ vanishes for $k \ne k_0$ and ρ is assumed bounded. Let us focus then on the first term, where we upper bound $P(T_{[P]_K})$ by unity:

$$\begin{aligned}
&\sum_{a^{i-1}} Q^*(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&= \sum_{a^{i-1} \in T_{Q^*}(\delta)} Q^*(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) + \\
&\quad \sum_{a^{i-1} \in T^c_{Q^*}(\delta)} Q^*(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})). \quad (7)
\end{aligned}$$

Once again, the second term vanishes as it pertains to a–typical sequences. As for the first term, we have:

$$\begin{aligned}
&\sum_{a^{i-1} \in T_{Q^*}(\delta)} Q^*(a^{i-1})\rho(P(\cdot), Q(\cdot|a^{i-1})) \\
&\le \sum_{a^{i-1} \in T_{Q^*}(\delta)} Q^*(a^{i-1})[\rho(P(\cdot), Q^*(\cdot)) + \mu_{m,K}] \\
&\le \sum_{a^{i-1} \in T_{Q^*}(\delta)} Q^*(a^{i-1})[\rho([P]_K(\cdot), Q^*(\cdot)) + \delta_K + \mu_{m,K}] \\
&\le \sum_{a^{i-1} \in T_{Q^*(\delta)}} Q^*(a^{i-1})(D + \delta_K + \mu_{m,K}) \\
&\le D + \delta_K + \mu_{m,K}. \quad (8)
\end{aligned}$$

Finally, we should add to the distance yet another term that is proportional to ϵ to account for all $i < \epsilon n$. This completes the proof of Theorem 2.

IV. Extension to Markov Sources

Theorem 1 and 2 can be extended to the Markov case, but this requires some more care. We next briefly review how this extension can be carried out for first–order Markov sources (further extension to higher orders is straightforward).

For simplicity, let us assume that Y^n is required to be stationary, which is a reasonable assumption when the input is stationary. We will also assume now that ρ is convex in Q. Let us now define

$$\begin{aligned}
\phi(D) = \quad & \max\{H(Y_1|Y_0) : \ \text{dist}\{Y_0\} = \text{dist}\{Y_1\}, \\
& \sum_a Q(a)\rho(P(\cdot|a), Q(\cdot|a)) \le D\}, \quad (9)
\end{aligned}$$

where $H(Y_1|Y_0)$ is the conditional entropy of Y_1 given Y_0 under the first–order Markov probability measure Q, and the maximization is over the transition probabilities $\{Q(b|a), \ a, b \in \mathcal{A}\}$, subject to the constraints that the unconditional marginal distributions, $\{Q(a), \ a \in \mathcal{A}\}$, of Y_0 and Y_1 are the same and the weighted distance constraint between the transition probability distributions $\{Q(\cdot|a)\}$ and $\{P(\cdot|a)\}$ is maintained. Also, let

$$\begin{aligned}
\phi(D; Q_0) = \quad & \max\{H(Y_1|Y_0) : \ \text{dist}\{Y_0\} = \text{dist}\{Y_1\} = Q_0, \\
& \sum_a Q_0(a)\rho(P(\cdot|a), Q(\cdot|a)) \le D\}, \quad (10)
\end{aligned}$$

and observe that for a given Q_0, $\phi(\cdot; Q_0)$ is concave (due to the convexity of ρ in Q). Then, for every $i = 2, \ldots, n$, we have

$$\begin{aligned}
D_i &\stackrel{\Delta}{=} \sum_{a^{i-1}} Q(a^{i-1})\rho(P(\cdot|a_{i-1}), Q(\cdot|a^{i-1})) \\
&= \sum_{a_{i-1}} Q(a_{i-1}) \sum_{a^{i-2}} Q(a^{i-2}|a_{i-1})\rho(P(\cdot|a_{i-1}), \\
&\quad Q(\cdot|a_{i-1}, a^{i-2})) \\
&\ge \sum_{a_{i-1}} Q(a_{i-1}) \cdot \rho(P(\cdot|a_{i-1}), \\
&\quad \sum_{a^{i-2}} Q(a^{i-2}|a_{i-1})Q(\cdot|a_{i-1}, a^{i-2})) \\
&= \sum_{a_{i-1}} Q(a_{i-1})\rho(P(\cdot|a_{i-1}), Q(\cdot|a_{i-1})) \stackrel{\Delta}{=} D'_i, \quad (11)
\end{aligned}$$

where the inequality follows from the assumed convexity of ρ. Thus, for any simulation scheme with a given marginal Q_0 of each Y_i, we have

$$\begin{aligned}
H(Y^n|X^m) &\le \sum_i H(Y_i|Y_{i-1}) \\
&\le \sum_i \phi(D'_i; Q_0) \\
&\le n\phi\left(\frac{1}{n}\sum_i D'_i; Q_0\right) \\
&\le n\phi(D; Q_0) \le n\phi(D). \quad (12)
\end{aligned}$$

The achievability scheme is constructed and analyzed in the same spirit as in Theorem 2 except that the memoryless structure is replaced by the Markov one.

V. The $\bar{\rho}$ Distance Measure

A related result is now developed for the $\bar{\rho}$ distance measure considered in [8] and [5], where distances between probability measures are induced by distortion measures between sequences of random variables. In this section, we are back to the memoryless case, and the results do not seem to lend themselves easily to extensions to sources with memory.

Let $\rho : \mathcal{A}^2 \to \mathbb{R}^+$ be a given single–letter distortion measure, and consider the Ornstein $\bar{\rho}$ distance, $\bar{\rho}(P, Q)$, between two measures P and Q of n–vectors in $\mathcal{A}^n$, i.e., the minimum of $\frac{1}{n}\sum_{i=1}^n E\rho(\tilde{X}_i, Y_i)$ across all joint distributions of $(\tilde{X}^n, Y^n)$ for which the marginal of $\tilde{X}^n$ is P and the

marginal of Y^n is Q.[3] Thus, loosely speaking, the $\bar{\rho}$ distance gives the best explanation of $Y^n \sim Q$ as a distorted version of $\tilde{X}^n \sim P$ via some channel. For a given distortion level D, we will allow the probability law Q of Y^n to be at $\bar{\rho}$ distance at most D from Q, i.e., $\bar{\rho}(P,Q) \leq D$.

In view of the above, consider the function

$$\Gamma_{n,m}(D) = \max\left\{\frac{1}{n}H(Y^n|X^m): \quad \bar{\rho}(P,Q) \leq D\right\}, \quad (13)$$

where, again, Q is understood as the probability measure that governs Y^n and P is the one that governs X^m. Next, define the single–letter function:

$$\gamma(D) = \max\{H(Y): \ E\rho(X,Y) \leq D\} \quad (14)$$

where $X \sim P$ and the maximization is across conditional distributions $\{W(y|x), \ x,y \in \mathcal{A}\}$ that satisfy the distortion constraint. It is easy to see that $\gamma(\cdot)$ is concave (simply because the entropy is concave).

For example, if P is binary with parameter $p < 1/2$, and ρ is the Hamming distortion measure, then denoting the binary entropy function by $h_2(t)$, $t \in [0,1]$, we have $\gamma(D) = h_2(p + D)$ for $D < 1/2 - p$ and $\gamma(D) = 1$ otherwise.

Our converse theorem asserts that $\gamma(D)$ is an upper bound to the per–symbol conditional entropy.

Theorem 3: (Converse): For all n and m, $\Gamma_{n,m}(D) \leq \gamma(D)$.

Proof. Given a simulation scheme that satisfies the $\bar{\rho}$ constraint, then by definition, there must exist a random vector $\tilde{X}^n \sim P$ such that $\frac{1}{n}\sum_{i=1}^n E\rho(\tilde{X}_i, Y_i) \leq D$. Thus,

$$\begin{aligned} H(Y^n|X^m) &\leq \sum_{i=1}^n H(Y_i) \\ &\leq \sum_{i=1}^n \gamma(E\rho(\tilde{X}_i, Y_i)) \\ &\leq n\gamma\left(\frac{1}{n}\sum_{i=1}^n E\rho(\tilde{X}_i, Y_i)\right) \leq n\gamma(D), \end{aligned} \quad (15)$$

where the first inequality is because conditioning reduces entropy, the second is by definition of $\gamma(\cdot)$, the third is due to the concavity of $\gamma(\cdot)$, and the fourth is due to its monotonicity and the aforementioned distortion constraint. This completes the proof of Theorem 3.

Theorem 4: (Direct): For all $m \geq n$,

$$\Gamma_{n,m}(D) \geq \gamma(D) - \epsilon_n,$$

where ϵ_n tends to zero as n grows without bound.

Sketch of Proof. If $m > n$, we will ignore the training samples $X_{n+1}, \ldots, X_m$, and so, reduce m to the value of n. Thus, from this point, we will assume $m = n$ and denote both integers by n. While $\gamma(D)$ depends on P, we next show that it is universally asymptotically achievable for large n. For a given P, let $Q = f(P)$ denote the output marginal induced by P and by the channel W that attains $\gamma(D)$. For a given training sequence x^n, let P_{x^n} denote the empirical distribution, and let $Q_n = [f(P_{x^n})]_n$, where the operation $[\cdot]_n$ means quantization of a given probability distribution to the nearest rational distribution with denominator n. The proposed simulation scheme will simply draw Y^n uniformly from the type class corresponding to Q_n (using the key U^ℓ for this random selection). Since R is assumed larger than $\gamma(D)$, the randomness of U^ℓ will suffice to implement a uniform distribution within the type class of Q_n, with high probability [4]. We now have to show that: (i) the output distribution of Y^n is (essentially) within $\bar{\rho}$–distance D from P, and (ii) performance is close to $\gamma(D)$ for large enough n. As for (i), consider a random vector $\tilde{X}^n$ drawn from P and let $W(Y^n|\tilde{X}^n)$ assign a uniform distribution on the conditional type class associated with the (single–letter) channel that achieves $\gamma(D)$. The uniform distribution within T_{x^n} induces a uniform distribution within the type class of Q_n, and at the same time, the distortion constraint is maintained by joint typicality. As for (ii), we have:

$$\begin{aligned} H(Y^n|X^n) &= \boldsymbol{E}\{\log|T([f(P_{X^n})]_n)|\} \\ &= n\boldsymbol{E}\{H([f(P_{X^n})]_n)\} - O(\log n) \\ &= n[\gamma(D) - \epsilon_n] \end{aligned} \quad (16)$$

where the last passage is due to the law of large numbers, the continuity of f, the vanishing effect of the operation $[\cdot]_n$, and the fact that $H(f(P)) = \gamma(D)$.

ACKNOWLEDGEMENT

This work was done while N. Merhav was visiting Hewlett–Packard Laboratories in the Summer of 2005.

REFERENCES

[1] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. IT–39, no. 3, pp. 752–772, May 1993.

[2] N. Merhav, "Achievable key rates for universal simulation of random data with respect to a set of statistical tests," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 21–30, January 2004.

[3] N. Merhav, G. Seroussi, and M. J. Weinberger, "Universal delay–limited simulation," *Proc. ISIT 2005*, pp. 765–769, Adelaide, Australia, September 2005.

[4] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 5–20, January 2004.

[5] N. Merhav and M. J. Weinberger, "Addendum to "On universal simulation of information sources using training data"," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3381–3383, September 2005.

[6] G. Seroussi, "On universal types," *Proc. of 2004 IEEE Intern'l Symp. on Inform. Theory (ISIT'04)*, p. 223, Chicago, USA, June/July 2004.

[7] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 3, pp. 634–646, May 1994.

[8] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, January 1996.

[9] K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Separation of random number generation and resolvability," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2237–2241, September 2000.

[3] We are deliberately denoting here the random vector corresponding to P by $\tilde{X}^n$, because it may not coincide with the training sequence although both are goverened by P.